First Steps in Building a Verb Valency Lexicon for Romanian

Ana-Maria Barbu

Romanian Academy, Institute of Linguistics
13 Calea 13 Sepetembrie, 050711, Bucharest, Romania
anabarbu@unibuc.eu

Abstract. This paper presents some steps in manually building a verb valency lexicon for Romanian. We refer to some major previous works by focusing on their information representation. We select that information for different stages of our project and we show the conceptual problems encountered during the first phase. Finally we present the gradually building procedure of the lexicon and we exemplify the manner in which the information is represented in a lexicon entry.

Key words: valency, verb, semantic preferences, sense disambiguation, Romanian

1 Introduction

The goal of our project is to build a lexicon of the verb valencies for Romanian. Valences are sets of elements required by a predicate. Such a valuable resource does not exist yet for this language. Unfortunately, we do not dispose of a large enough corpus of Romanian texts, or a syntactically annotated one, as a starting point. In these conditions, we have to build this lexicon in several steps, by starting from the printed dictionaries, going to the corpus and coming back to improve the lexicon.

The main concern for achieving this task is the nature of the information we have to gradually gather in entry descriptions in order to use this resource first for syntactical dependency annotation, then for sense disambiguation, and finally for knowledge representation. At the first glance there is a rich literature on (automated or not) building subcategorisation or semantic frames from corpora. Nevertheless there are few works dealing with our problem which is building such a resource from scratch. Therefore, we insist on the conceptual part of the project by highlighting aspects which seem not to be referred in the literature.

In the first section of this paper, we review previous researches on this topic. The main section is devoted to the analysis of the points we take over from these approaches and the information necessary to be encoded. We also discuss the problems we encountered. Another section describes the present stage of our lexicon: the building steps and the entry structure. Finally, we show the further steps in aiming at our goal.

2 Previous work analysis

In this section we briefly review the previous approaches we used as a starting point for our project, namely FrameNet, VerbNet, VALLEX and CPA. We especially pay attention to the information encoded in each studied framework. Then we analyze what information is proper to take over for our lexicon, at this stage and in the future, and also whether some other information is needed.

FrameNet. We confine ourselves to point out only some of the FrameNet project characteristics described in [1].

- Parts in British National Corpus (BNC) are annotated with frame semantic information.
- Semantic frames mainly use five types of information: *domains*: Body, Cognition, etc., *frame elements* (FE) expressed by semantic roles: Buyer, Seller, etc., *phrase types* (PT): Noun Phrase, Verb Phrase etc., *grammatical functions* (GF): External Argument, Complement etc., and cases of *nonexpressed frame elements*: Null Instantiation of Indefinite (INI), Definite (DNI) and Constructionally licensed (CNI) types.
- The phrase types and grammatical functions are automatically inferred from BNC, which is a syntactically annotated corpus.

An attempt to use FrameNet for building a Romanian valency lexicon has been done within the project described in [2]. This supposes a parallel Romanian-English FrameNet using annotation import. The XML files of the annotated English sentences are aligned word-by-word with its corresponding Romanian translation, then it is automatically created a set of XML files containing a corpus of FE-annotated sentences for Romanian language. In our opinion this procedure suffers from the following drawbacks. First, it needs a translation of English texts or a huge amount of Romanian-English parallel texts which, so far, are lacking. Second, the translation introduces an artificial intervention of translators by choosing preferred words and structures. Third, the procedure inherits all the flaws of each automatic task it uses. In general, we plead for avoiding the translation of foreign resources for Romanian, especially those which are language-specific to a large extent and the final quality of which is as important as a valency lexicon.

VerbNet. VerbNet project described in [3] has the following characteristics.

- It use a variant of Levin verb classes to systematically construct lexical entries.
- Each verb refers to a set of classes corresponding to the different senses of the verb (e.g. 'run' refers to *Manner of Motion* class for 'John runs', and to *Meander* class for 'The street runs through the district').
- For each verb sense specific selectional restrictions and semantic characteristics are added, if they can not be captured by the class membership.

- Each verbal class lists the thematic roles and the selectional restrictions for each argument in each frame, as well as the syntactic frames corresponding to licensed constructions, which are linked to syntactic trees in Lexicalized Tree Adjoining Grammar representation. Besides, each member of a verbal class is mapped to the appropriate WordNet synset.
- Each frame also includes semantic predicates describing the participants during various stages of the event (namely *during*(E), *end*(E) and *result*(E)).

VALLEX. The Valency Lexicon of Czech Verbs (VALLEX 2.0) uses the information described in [4] and refers to the following aspects.

- A dictionary entry represents a lexeme structured on lexical forms and lexical units.
- A lexical form specifies the 'base' infinitive form (eventually with its variants), its morphological aspect (imperfective, perfective, etc.), aspectual counterparts, reflexive particle, if it is the case, and numbers for homographs, if they exist.
- Lexical units correspond to the meanings of the lexeme. Each lexical unit displays obligatory information: valency frame, gloss (a synonym or a paraphrase) and example, as well as optional information: flag for idiom, information on control, possible type(s) of reflexive constructions, possible type(s) of reciprocal constructions and affiliation to a syntactico-semantic class.
- Valency frames contain the following types of information assigned to frame slots:
  - Functors (Actor, Addressee etc.) structured on inner participants, quasi-valency complementation and free modification.
  - Morphemic forms (case, prepositional case, infinitive construction etc.)
  - Types of complementations (obligatory or optional for inner participants and quasi-valency complementations, and obligatory or typical for free modifications).
  - Marks for slot expansion.

CPA. "Corpus Pattern Analysis" (CPA) project, see [5] and [6], has the goal to extract, from BNC, all normal patterns of use of verbs and to assign a meaning to each pattern. The relationship between a pattern and its meaning seems to be of type one-to-one, that is, unambiguous. An entry of CPA lexicon displays the following characteristics.

- Each valency contains the following types of information:
  - A pattern described in terms of semantic values (*semantic types* and *semantic roles*), lexical sets, words or/and morpho-syntactic characteristics (e.g. [[Person]] grasp {[[Abstract]]—[N-clause]}).
  - One or more implicatures describing the meaning in terms of synonyms or paraphrases of the entry verb, and semantic values for its arguments (e.g. [[Person=Cognitive]] understands {[[Abstract=Concept]]—[N-clause]}).
  - Lexical alternation (e.g. [[Person]] <–> {hand, finger}).
  - Lexical set (e.g. [[Oportunity<Abstract]]: opportunity, chance, offer, moment...).

- Clues, that is, specific elements (words, phrases) usually used in the context of the respective meaning able to disambiguate that meaning.
- Comment, especially for marking idioms.

- A *semantic type* is a class to which a term is assigned (e.g. Person, BodyPart).
- A *semantic role* is the context-specific role of the related semantic type. For instance, a semantic type Person can play in a health context the semantic role Doctor: [[Person=Doctor]] or Patient: [[Person=Patient]].

3 Analysis of required information

We are now trying to analyze the types of information used in the above mentioned approaches, in order to select them in accordance with the needs of the different stages of our project. We also discuss some problems they raised in our representation. Information generally used is of the following kinds.

Phrase types. Information concerning the morpho-syntactic nature of valency elements is present in every approach either extracted from corpora (CPA, FrameNet) or built manually (VerbNet, VALLEX). We also need to represent it in our lexicon especially because there is no syntactic frames lexicon for Romanian and such a resource is basic for almost any task in NLP.

It is worth mentioning here an aspect concerning free alternation. Alternation preserving the meaning can be done either locally, for instance, one of type AdvP/PP: *He behaves well / in snobbish way*, or by restructuring the whole frame: *Imi place de Ion* ('Me like DE John') / *Eu il plac pe Ion* ('I him like PE John'). The local alternation can be regular, such as AdvP/PP, or verb-specific (NP/PP): *Ion priveste ceva / la ceva* ('John watches something / at something'). We tried to capture all these cases in our representation.

Lemma characteristics. Even if there is an obvious relationship between some morphological characteristics of a lemma and its valency frames, as it is also pointed out in [7], only VALLEX takes this into account. Some meanings can be appropriate, for instance, only for a verb at third person singular form (e.g. with impersonal use), or only for its negated form. On the other hand, only some meanings allow certain diathesis alternations and they have to be marked as such. Therefore we have assigned a slot for describing lemma characteristics even at this stage of our project.

Grammatical functions. Only FrameNet provides explicit information about grammatical functions of the frame elements. A Romanian approach on valence frames [8] also proposes a description in terms of grammatical functions based on the inventory in traditional Romanian grammar. However, we set aside this kind of information because it is too dependent on the theoretical background and because it is not reflected, in any way, in what we can extract from corpora. Besides, for a rich inflected language like Romanian functions can be often inferred from grammatical cases of arguments.

Selectional preferences. Surprisingly, this information is only captured in VerbNet and in CPA (as semantic types), despite its importance, in our opinion,

for distinguishing valencies and for word sense disambiguation. Therefore we pay special attention to them.

For our description we set up the following criterion: *two valency frames are distinct if their elements display different selectional preferences*. For instance, the intransitive usage of the Romanian verb *a toca* ('to hash') roughly means to make repeated short noises, to knock, to rattle. This valency (of intransitive form) has three different meanings depending on the semantic type of the subject. If the subject is a Person, the verb means 'to hammer on a board (for asking people to church)', if the subject is a Gun, the verb means that the (machine) gun is fired and if it is a Stork, the meaning is that the stork rattles its beak. Therefore we actually record three valency frames: NP[Person], NP[Gun], NP[Stork], corresponding to the three meanings, instead of one general frame such as NP[Concrete]. Note that the intransitive form of the verb *a toca* cannot have other subject type than a person, a machine gun or a stork with the corresponding meanings. For instance, in a sentence such as 'The house shook and the doors and windows rattled' Romanian does not translate 'to rattle' by *a toca*, even if the common feature of the three above mentioned meanings is 'to make repeated short noises'.

A consequence of our criterion (also mentioned above) is that, if given two meanings (registered in printed dictionaries) about which we cannot specify different semantic types for the elements of their corresponding valencies, then we assign both meanings to one and the same valency frame. For instance, the verb *a omori* ('to kill') means either 'to cause to die' or 'to torture' someone. For both senses, the subject can be practically anything and the object has to be Animate. At this stage of our project, we can not distinguish two different valency frames for the two meanings, but a future solution could be to implement semantic predicates, in VerbNet fashion, for marking the result aspect of the first meaning and the durative one of the second or/and to introduce Clues, in CPA style, because it is obvious that only external elements in context can disambiguate them.

Using semantic types for disambiguating meanings raises the following problem, as well. If there are two semantic types, one more general than the other, the problem is to determine to what extent the particular semantic type defines a new meaning. So for example, the verb *a lepada* has the meaning '[Person] sheds [Concrete]'. If Concrete=Clothes, its meaning is '[Person] sheds/takes off [Clothes]'. The question is, on the one hand, to what extent the meaning changed so that two valency frames are needed, i.e. '[Person] sheds [Concrete]' and '[Person] sheds [Clothes]'. On the other hand, if there are two meanings, can one assume that Clothes do not ever participate at the first meaning, so that no ambiguity exists? So far, in cases of subordinated semantics types without major difference of meanings we retain only the more general valency frame.

We do not use semantic types in a pre-determined inventory. We implement them in as much as they are needed. A problem we encounter is to establish the most general semantic type, able to include all the real situations a verb can imply. For instance, a verb like *a exprima* ('to express') (or any other denotation

verb) can have as its subject a person, a text, an image, an attitude, a situation, a sign etc. Which is the appropriate semantic type for all these cases? In these situations we preferred not to specify any semantic type on the subject. We think not even *lexical set* in CPA could solve the problem, but at the most we can choose a semantic type like Entity.

As it is mentioned in [6], sometimes it is worth putting down the axiological aspect of an element. *A oferi* ('to offer') has a complement which normally has a positive connotation. One offers good things, unlike 'to give' which is neutral. So, axiological semantic types have also to be introduced in our inventory. Besides, it turns out that a valency element may be described by more than one semantic restriction. It is likely that *sets* of semantic types are needed for describing one argument.

Obligatory vs. optional elements. This kind of information is explicit in FrameNet and VALLEX and implicit in VerbNet and CPA, where it is covered by enumerating all possible valency frames. We have chosen the method adopted for Czech lexicon in that we mark each argument as obligatory or optional. For a manually built lexicon this way is more efficient. However there is a problem here too. For a verb like *a se desfasura* ('to take place', 'to proceed') arguments indicating the place, the time and/or the manner in which an event can take place are needed. The problem is how to express the fact that any of these arguments can miss but not all of them. It is not clear how the mentioned approaches grasp, on the one hand, the fact that in the sentences: 'The show takes place tomorrow / here / somehow', 'The show takes place here tomorrow ' etc. the verb 'take place' has the same meaning and, on the other hand, 'The show takes place' is ungrammatical. In other words, we need a mean of marking that, in this case, any argument is optional but at least one is obligatory.

Modifiers. FrameNet, VALLEX and CPA include modifiers in valency structures. At the first stage, we confine ourselves to represent the 'minimal' valencies because of the lack of accessible corpus evidence. Furthermore, this kind of information has to be added in order to reflect the typical use of verbs and to distinguish meanings more precisely (see Clues in CPA).

Meanings. Sense descriptions are made by means of synonyms and paraphrases (VALLEX), WordNet synsets (VerbNet) or 'primary implicature' (CPA). We join VALLEX by using synonyms and paraphrases for describing one or more meanings of a valency. In general, we follow the meanings displayed in a verb entry of printed dictionaries. However, there are cases in which we have to split one dictionary definition because it corresponds to more than one valency structure and meaning. On the other hand, there could be, for instance, two definitions for which we are not able to say precisely what distinguishes the another and thus we describe one valency frame with two meanings. For instance, with the means we use to describe valencies at this stage we are not able to assign different valency frames to the verb *a trai* 'to live' for the sentences *Ion isi traieste tineretea (intens)* 'John lives his youth (intensively)' and *Spectatorii au trait momentul (cu entuziasm)* 'The public lived the moment (enthusiastically)' (see Fig.1 below) even if they represent two different meanings in the printed dictionary.

Semantic roles. This information is referred in all studied approaches but always in different ways. The common point is that they reflect the role which an argument plays *in context* unlike the semantic types or selectional preferences which reflect lexical features of a given lexeme, able to be organized into a semantic ontology. We postponed specifying semantic roles in our valencies description because there is not an unanimously accepted inventory, so that we have to reflect on criteria for building or adopting it and because semantic roles are less informative for our immediate purposes.

Verb classes. Classes are useful for generalizations and systematic work. They are extensively used in VerbNet, but also in VALLEX. An approach which mainly distributes valencies in classes is [10], as well. We have postponed any classification for the moment when we get a significant inventory of verb valencies.

Semantic predicates. This information described in VerbNet or [9] is very important for knowledge representation. However it goes beyond the immediate aims of our project.

4 Building steps and results

Building the lexicon means four main steps.

I. Describing subcategorization frames which provide morpho-syntactic information on verb arguments and their 'primary' semantic types.

II. Extended search on corpus, by using the lexicon obtained at the previous step, for adding information on modifiers and for refining semantic types.

III. Adding information about control and passivization and implementing rules for expanding structural regularities and alternations in valence representation.

IV. Adding information about semantic roles and semantic predicates.

The tasks of the first stage of the project have been fulfilled. First, we have chosen about 3000 verbs from the Romanian core lexicon. The main resource for getting the senses of the verbs was a Romanian Explanatory Dictionary (DEX). The information in the printed dictionary was confronted with Romanian texts on Internet and a corpus of about 14 millions words from newspapers. This task aimed at actualizing senses of verbs, getting primary semantic types and examples for valency lexicon. Entries were built manually, during about three years, by a team of five linguists. They followed the meta-language described in [11], where the significance of the formal representation and the grammatical characteristics appropriate for describing verb valencies for Romanian are fully detailed.

The result of the first phase is a lexicon which, this year, reaches a number of about 3000 verbs, in a text format and XML format. Fig. 1 shows three (from ten) argument structures of the verb *a trai* ('to live') in text format and the corresponding XML representation of the second argument structure. The XML format is automatically obtained from the text description which is meant to

follow the meta-language strictly. Any enhancement on lexicon is done on the text format, which is afterwards translated into the XML one.

1. Entry in text format	2. Entry in XML format
***a trăi** 'to live'* Argument structures: 1. NP[*nom*, +animate] Senses: ▪ to live: *Victima trăieşte.* 'The victim is alive'. 2. NP[*nom*, +animate] NP[*ac*, +period] (AdvP[manner] or PP[*la/cu*, -]) Meanings: ▪ to spend: *Ion îşi trăieşte tinereţea (intens / la maxim)*. 'John lives his youth (intensively / to the maximum)'. ▪ to feel intensively: Spectatorii au trăit momentul (cu entuziasm). 'The public lived the moment (enthusiastically)'. 3. NP[*nom*, +animate] PP[*pentru*, +goal] Meanings: ▪ to devote his/her life: Femeia trăieşte pentru răzbunarea soţului. 'The woman lives for avenging her husband'.	<entry> <verb msr="">trăi</verb> <str-arg. nr="2"> <arg nr="1" caz="nom", restr-sem="+animat">GN</arg> <arg nr="2" caz="ac", restr-sem="+perioadă">GN</arg> <arg nr="3"> <arg nr="3.1" tip="mod" restr-sem=""> GAdv</arg> <arg nr="3.2" tip="" prep="la, cu" restr-sem=""> GP</arg> </arg> <sensGrup> <sens nr.1> <sin>a petrece</sin> <eg> Ion îşi trăieşte tinereţea (intens / la maxim);</eg> </sens> <sens nr.2> <par>a resimţi intens</par> <eg> Spectatorii au trăit momentul (cu entuziasm)</eg> </sens> </sensGrup> </str-arg> </verb> </entry>

Fig. 1. Examples of an entry in Romanian Verb Valency Lexicon

5 Conclusions

This paper approaches the building of a verb valency lexicon for Romanian. It mainly presents some conceptual problems encountered while working on it and not referred in previous works. On the other hand, the paper shows that one can manually develop such a valuable resource based on deep linguistic insights, quick enough, iteratively, first meant for basic computational purposes, then for complex ones, despite the "mirage" of the automatic processing present in the literature.

The result of the first step of the project is a lexicon covering the valency frames of about 3000 verbs, which can be used in shallow and deep parsing and in tasks of word sense disambiguation. Refining the semantic preferences and adding new information of knowledge representation are targets for new steps.

Acknowledgments. The research reported in this paper is done in the framework of the grant no. 1156/A, founded by the National University Research Counsel of Romania (CNCSIS).

References

1. Johnson, C., Fillmore, C.: The FrameNet tagset for frame-semantic and syntactic coding of predicate-argument structure. In 1st Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL 2000), pp. 56-62, Seattle WA (2000)
2. Trandabat, D.: Semantic Frames in Romanian Natural Language Processing. In Proceedings of the North American Chepter of the Association for Computational Linguistics NAACL-HLT 2007, Companion Volume: Doctoral Consortium, Ed. Association for Computational Linguistics, April 2007, pp. 29-32, Rochester, New York, USA, (2007)
3. Kipper, K., Dang, H. T., Palmer, M.: Class-based Construction of a Verb Lexicon, AAAII-2000 Seventeenth National Conference on Artificial Intelligence, Austin, TX, (2000)
4. Zabokrtsky, Z., Lopatkova, M.: Valency Information in VALLEX 2.0: Logical Structure of the Lexicon. The Prague Bulletin of Mathematical Linguistics No. 87, 41-60, (2007)
5. Hanks, P., Pustejovsky J.: A Pattern Dictionary for Natural Language Processing. Revue franaise de linguistique applique, X-2, 63-82 (2005).
6. Hanks, P.: The Organization of the Lexicon: Semantic Types and Lexical Sets, http://www.cs.cas.cz/semweb/download/06-11-hanks.doc (2006).
7. Przepiorkowski, A.: Towards the Design of a Syntactico-Semantic Lexicon for Polish. In New Trends in Intelligent Information Pocessing and Web Mining, Zakopane, Springer Verlag, (2004).
8. Serbanescu A. (1994) Pentru un dictionar sintactic al verbelor romnesti, Studii si Cercetari Lingvistice, XLV, nr.3-4, 133-150, Bucuresti, (1994).
9. Pustejovsky, J.: The Generative Lexicon, The MIT Press, Cambridge, Massachusetts, London, England, (2001).
10. Leclere, C.: The Lexicon-Grammar of French Verbs: a syntactic database. In: Kawaguchi, Y., *et alii* (eds.) Linguistic Informatics - State of the Art and the Future, Tokyo University of Foreign Studies, UBLI 1, pp. 29-45. Benjamins, Amsterdam / Philadelphia (2003)
11. Barbu, A.M., Ionescu, E.: Designing a Valence Dictionary for Romanian. In International Conference Recent Advances in Natural Language Processing (RANLP), p.41-45, Borovets, Bulgaria, (2007).